日英のコーパスを用いたプロソディ研究*

北原 真冬（上智大学外国語学部）
mafuyu@sophia.ac.jp

1. はじめに

本発表の射程は(1)英語母語話者，(2)日本語母語話者，そして(3)日本語を第一言語とする英語学習者，のプロソディ研究である．実は(3)を対象にしようとすると，必然的に(1)をターゲット，(2)をベースラインとして押さえておかねばならない．研究プロジェクトの規模を無闇に拡大しないために，特に(1), (2)についてコーパスを用いた下調べは必須である．単語頻度・親密度データベースとして(1)に Hoosier Mental Lexicon (HML), (2)に日本語の語彙特性(Psylex)，自然発話コーパスとして(1)に Buckeye Corpus，(2)に Corpus of Spoken Japanese (CSJ)を用いた研究のデザインを粗描する．

2. 単語頻度・親密度データベースの利用

音声の産出や知覚について実在する単語を用いて実験をする場合，実験の焦点となる変数以外についてはなるべく平等な条件を揃えたい．例えば，単語アクセントの有無と句頭のピッチの関わりを焦点とする実験を組むには，無アクセントの単語群と有アクセントの単語群を用意することになる．この時，二つの単語群の親密度は一定の基準範囲に収まるように揃えることが望ましい．どちらかの単語群になじみのない単語が含まれていると，結果にバイアスが入ってしまう恐れがある．また，各単語の近傍(neighborhood)の密度や頻度のバランスも考慮することが望ましい．以下では，それらの統制に必要な英語と日本語の単語親密度データベースを概観する．

2.1. Hoosier Mental Lexicon (HML)

HML (Nusbaum, et al. 1984)はインディアナ大学心理学部において開発された単語親密度データベースである．Webster Pocket Dictionary (1964)の見出し語約 2 万項目について，参加者が 7 段階の尺度で親密度を評定したデータが, 単語の音韻表記等と共に収録されている．

表 1. HML 所収データのサンプル

正書法	音韻表記	頻度	親密度	内容/機能語	品詞
a	x	23237	7.00000	f	=nr
aback	x-b'@k	2	5.08333	c	a
abacus	'@<bx-kxs	0	5.16667	c	n

* 本発表は国立国語研究所共同研究プロジェクト「対照言語学の観点から見た日本語の音声と文法」の一部である．また，田嶋圭一氏(法政大学)，米山聖子氏(大東文化大学)との共同研究の成果を含む．関係各位に感謝する．

2.2. Psylex

日本語の語彙特性(天野・近藤, 1999)は NTT コミュニケーション科学基礎研究所において開発された単語親密度を中心としたデータベースである．新明解国語辞典第四版の見出し語約 7 万項目について，HML と同様の 7 段階尺度を用いて，音声だけでなく表記についての親密度と音韻表記等を収録している．他にも，朝日新聞約 14 年分のデータにおける単語の出現頻度，単語の心像性についての評定値，3 万語の増補も併せた総計 9 巻に及ぶデータが三省堂から出版された. ID 番号によって 1,3,7 巻から情報を集約したサンプルは以下の通りである．

表 2. Psylex 所収データのサンプル

ID	カナ	表記	長さ	アクセント	ローマ字	頻度	AV 親密度	A 親密度	V 親密度
434930	ハシ	端	2	0	ha.shi	2440	5.125	5.906	5.156
434940	ハシ	嘴	2	1	ha.shi	21	2.219	5.500	2.938
434950	ハシ	箸	2	1	ha.shi	277	5.625	5.500	5.250
434960	ハシ	橋	2	2	ha.shi	3975	6.250	5.906	6.062

3. 自然発話コーパスの利用

3.1. Buckeye コーパス

Buckeye コーパスは米国オハイオ州立大学の Mark Pitt らをリーダーとするチームによって作成された自然発話コーパスである．オハイオ州コロンバス在住の成人男女 40 名を対象としたインタビュー形式で，およそ 30 万語分の発話が収録されている。音声データは，まず自動認識処理によってラベル付けが与えられたのち，訓練を受けた音声学者が修正した．

Buckeye コーパスは各話者の録音が数個のセッションに分けられており, sXXYYZ という形式のフォルダに各セッションが収められている．ここで XX は話者番号，YY はセッション番号，Z はセッション記号であり，例えば s0101a は話者番号 01，セッション 01a という意味である. 各フォルダは音声ファイル(sXXYYZ.wav), 書き起こしテキスト(sXXYYZ.txt), 及び音声解析ソフトウェア Praat のアノテーション(sXXYYZ.TextGrid)を含んでいる．

Praat のサウンドエディターで wav ファイルと TextGrid ファイルを開くと図 1 のような画面表示となる．1 番上のパネルは音声波形，2 番目はスペクトログラム，その下のテキストの入った枠は，単語層，分節音層，コメント層が順に並んでいる．

単語層は全て小文字を用いているが，基本的に英単語のスペルそのままであり，語末にセミコロンが付されている．また図 1 の冒頭部にあるように「<HES>itation=ためらい」など，いくつかのタグが用いられている．分節音層は DARPA phonetic alphabet に準拠した音声表記であり，IPA などの特殊フォントを用いていない．使用されているタグ，記号，コメントなどの一覧は付属するマニュアルに記載されている．

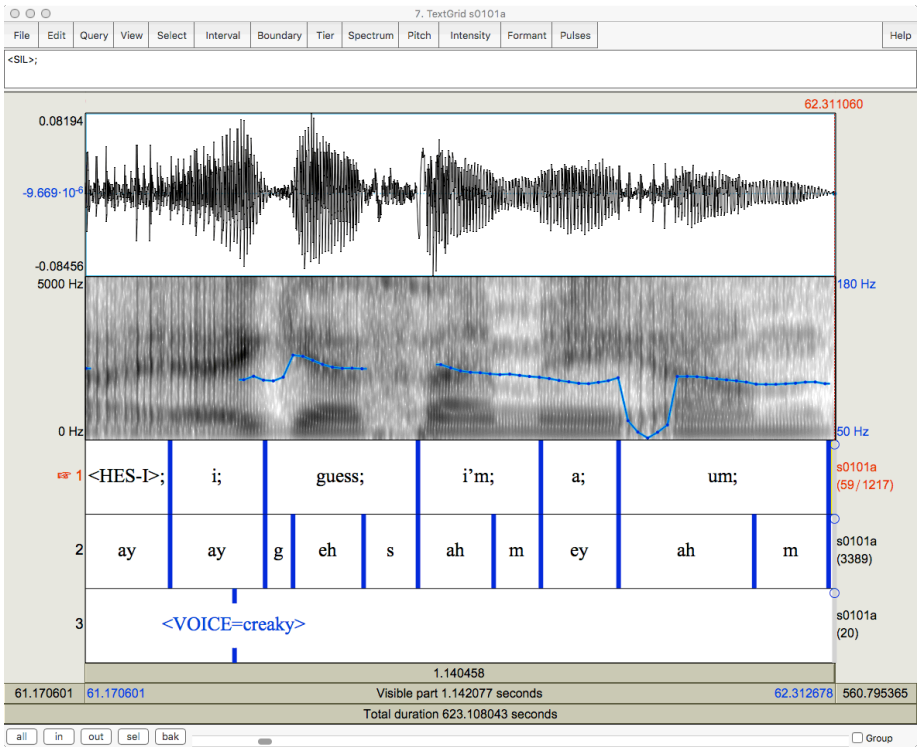

図 1: Buckeye コーパスを Praat で開いたところ

ここでは，Buckeye コーパスにおける歯茎弾音(alveolar flap)の生起についての調査を分析の具体例として示す．歯茎弾音は北米の英語において，/t/, /d/が強勢のある母音と次の母音に挟まれた時に生起する．例えば，*better, rider* のように語中で起こるだけでなく，*get up, need it* のように単語間でも起こる．Praat の TextGrid でデータが提供されているため，比較的簡単な Praat スクリプトによって例えば以下のようなデータを得ることができる．

表 3. Buckeye コーパスにおける検索のサンプル

単語	検索対象	持続時間(秒)
thirty;	dx	0.020144999999999413
created;	dx	0.026443999999997914

筆者を含む研究グループでは，これに基づき，帰国子女を含む様々なグループの日本人英語学習者の弾音化と比較することで，英語学習における非音素的な特徴の役割を解明しようとしている．その際，コーパスの探索は実験のベースラインを与えてくれるという意味で極めて重要である．いわゆる実験室発音(lab speech)ではない，自発的な会話の音声データを検索し，任意の音声現象について環境ごとの出現率が得ることで，より特定的な条件や狙いを定めた単語群についての発話産出実験や知覚実験をデザインできる．

3.2. 日本語話しことばコーパス

日本語話し言葉コーパス(Corpus of Spoken Japanese: CSJ)は，国立国語研究所，情報通信機構，東京工業大学の 3 者が共同で開発した，大規模な自発音声のデータと詳細なアノテーションを持つコーパスである．Buckeye コーパスに比べると，形態・統語情報，音声・韻律情報などが豊富に用意されている．ここでは北原・米山(2014)において扱った，母音の持続時間に関する後続子音の有声性の影響を一例として紹介する．

英語の母音が有声子音の前では無声子音の前よりも約 50%も持続時間が長いことはよく知られている．しかし，有声子音の前で母音が長くなるのは，調音的に見て自然な同化プロセスであり，日本語の発話においても英語ほどではないにしても，似たような傾向が見出せるかどうかは調査に値する．まず日本人乳幼児の発話データベース(Amano et al., 2008)を調べると，後続子音の有声性による母音長への影響が見られた．つまり，調音的に自然なプロセスが乳幼児の日本語発話においては現れる．ところが成人になると，後続子音の影響は弱まり，短母音の長さにほとんど差がない場合も多いことが CSJ の分析から分かった．ここでの CSJ の利用法は，Buckeye コーパスの場合と基本的には変わらず，コーパス内にあるアノテーションのみを用いている. Praat の TextGrid に対するスクリプトを用いれば，任意の分節音についての持続時間やその周辺のラベルを抽出することは容易である．

4. まとめと展望

本発表では主に 4 つのデータベース・コーパスを扱い，それらを元に日本語および英語のプロソディ分析に役立つ情報の取り出し方について触れた．これらは分析の基礎的作業だが，スクリプト言語によるプログラミングを含むため，一般には敷居が高く感じられることも多い．今後，なんとか敷居を下げる方策も探していきたいと考えている．

参考文献

天野成昭・近藤公久 (1999)『NTT データベースシリーズ:日本語の語彙特性』東京：三省堂

Amano, S., K. Tadahisa, and K. Kato (2008) "Development of NTT Infant Speech Database," *Technical report of IEICE*, 108(59), 29-34.

Boersma, P. and D. Weenink (2017) *Praat: doing phonetics by computer* [Computer program]. Version 6.0.29, retrieved 24 May 2017 from http://www.praat.org/

北原真冬・米山聖子 (2014) 「後続子音による母音長の変化：幼児・成人の日本語コーパス分析と成人の英語学習データ」*JELS* 31, 44-48.

国立国語研究所 (2006) 『日本語話し言葉コーパス』[pj.ninjal.ac.jp/corpus_center/csj/]

Nusbaum, H. C., D. B. Pisoni, and C. K. Davis (1984) "Sizing up the Hoosier Mental Lexicon," *Research on Speech Perception Progress Report* 10, 357-375. Bloomington: Indiana University

Pitt, M. A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier (2007) "Buckeye Corpus of Conversational Speech (2nd release)," [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor)